arXiv:1301.4157v1 [cs.LG] 17 Jan 2013

On the Product Rule for Classification Problems

Marcelo Cicconet
New York University
cicconet@gmail.com

Keywords: Supervised Learning; Classification; Product Rule.

Abstract: We discuss theoretical aspects of the product rule for classification problems in supervised machine learning for the case of combining classifiers. We show that (1) the product rule arises from the MAP classifier supposing equivalent priors and conditional independence given a class; (2) under some conditions, the product rule is equivalent to minimizing the sum of the squared distances to the respective centers of the classes related with different features, such distances being weighted by the spread of the classes; (3) observing some hypothesis, the product rule is equivalent to concatenating the vectors of features.

1 Introduction

With the advance of the Machine Learning field, and the discovery of many different techniques, the subject of *combining multiple learners* [2] eventually drove attention, in particular the problem of *combining classifiers*. Many different methods appeared, and soon they were compared in terms of efficiency in solving problems.

The *product rule* has been present in some of these works (e.g., [1, 7, 3, 6, 5, 4, 8]), in contexts ranging from the accuracy of the different combination rules to some analytical properties of the different methods.

In [3] it was shown that, in the context of handwritten digit recognition, the product rule performs better for combining linear classifiers. In general, however, the product rule does not stand out from competitors [6]. For the problem of combining audio and video signals in guitar-chord recognition, the product rule is better then the sum rule [5], but on the problem of identity verification using face and voice profiles, the sum rule wins [7].

On the theoretical realm, [1] shows that for problems with two classes, the sum and product rules are equivalent when using two classifiers and the sum of the estimates of the a posteriori probabilities is equal to one. In [7], the product rule is derived from the hypothesis of conditional statistical independence between different representations of the data. There are also some intuitive explanations for the choice of the product rule, as for instance the fact that the product ("END" operator) is preferred with respect to the sum rule ("OR" operator) because it enforces all qualities defined by the measures at once [9].

In this text, analytical properties of the product rule are further analyzed, in the contexts of two or more classifiers. We show that (1) the product rule arises from the MAP classifier supposing equivalent priors and conditional independence given a class; (2) under some conditions, the product rule is equivalent to minimizing the sum of the squared distances to the respective centers of the classes related with different features, such distances being weighted by the spread of the classes; (3) observing some hypothesis, the product rule is equivalent to concatenating the vectors of features.

Our work extends the current theoretical understanding of the product rule provided by Alexandre *et al* [1] and Kittler *et al* [7], as it was made in the direction of the sum rule by Li and Zong [8].

2 Theoretical Facts

Definition 1. *Let X,Y be (continuous) random variables corresponding to 2 distinct feature vectors, and C the (discrete) random variable corresponding to the class, whose output can be $c_1,...,c_K$. For any $Z \in \{X,Y\}$ and $k \in \{1,\dots,K\}$, let $p_{Z,k}$ be a function that outputs the* confidence *that the class is c_k considering that the features-variable is Z. Supposing that the features are $X = x$ and $Y = y$, the* product rule *for classification will assign $C = c_{\hat{k}}$ provided*

$$p_{X,\hat{k}}(x) \cdot p_{Y,\hat{k}}(y) = \max_{k=1,...,K} p_{X,k}(x) \cdot p_{Y,k}(y) .$$

In this definition and in the following results we are using, for simplicity, only two random variables, named X and Y. We could have used, instead, a set of N random variables, say $X^1, \ldots, X^N$, but that would unnecessarily overload the notation.

Definition 2. *Let (X,Y) be the random variable obtained by concatenating the features X and Y, and $p(\cdot|C=c_k)$ the density function for the variable (X,Y) conditioned to $C=c_k$. We will denote the value of this function at the point (x,y) by $p(X=x, Y=y|C=c_k)$. Let $P(C=c_k)$ be the* prior *probability that the class is $C=c_k$.*

Finally, let us define $p_{(X,Y),k}(x,y)$ as follows:

$$p_{(X,Y),k}(x,y) = p(X=x,Y=y|C=c_k) \cdot P(C=c_k) \;.$$

Given a sampled value $(X,Y)=(x,y)$, the MAP *(Maximum a Posteriori) classifier will assign $C=c_{\hat{k}}$ provided*

$$p_{(X,Y),\hat{k}}(x,y) = \max_{k=1,\ldots,K} p_{(X,Y),k}(x,y)$$

Fact 1. *When using the MAP classifier, the product rule arises under the hypothesis of (1) conditional independency given the class and (2) same prior probability for the classes.*

Proof. The MAP classifier is given by

$$p(X=x,Y=y|C=c_k) \cdot P(C=c_k) \;.$$

Now hypothesis 1 means

$$\begin{gathered} p(X=x,Y=y|C=c_k) = \\ = p(X=x|C=c_k) \cdot p(Y=y|C=c_k) \;, \end{gathered}$$

and hypothesis 2 implies that $P(C=c_{\tilde{k}}) = P(C=c_{\hat{k}})$ for all $\tilde{k}, \hat{k} = 1, \ldots, K$. Therefore

$$\begin{gathered} \max\nolimits_{k=1,\ldots,K} p_{(X,Y),k}(x,y) = \\ = \max\nolimits_{k=1,\ldots,K} p(X=x|C=c_k) \cdot p(Y=y|C=c_k) \;, \end{gathered}$$

which is the product rule (see definition 1) for $p_{X,k}(x) = p(X=x|C=c_k)$ and $p_{Y,k}(y) = p(Y=y|C=c_k)$.

□

Fact 2. *For each $Z \in \{X,Y\}$, let d_Z be the (finite) dimension of the variable Z, I_{d_Z} the identity matrix of dimensions $d_Z \times d_Z$, and $\Sigma_{Z,k} = \sigma_{Z,k}^2 I_{d_Z}$ (where $\sigma_{Z,k}$ is positive number). Also, for each $k = 1, \ldots, K$, let $\mu_{Z,k}$ be fixed points in $\mathbb{R}^{d_Z}$.*

Defining confidence functions (see definition 1)

$$p_{X,k}(x) = e^{-\frac{1}{2}(x-\mu_{X,k})^\top \Sigma_{X,k}^{-1}(x-\mu_{X,k})} \text{ , and} \qquad (1)$$

$$p_{Y,k}(y) = e^{-\frac{1}{2}(y-\mu_{Y,k})^\top \Sigma_{Y,k}^{-1}(y-\mu_{Y,k})} \;, \qquad (2)$$

the product rule is equivalent to

$$\min_{k=1,\ldots,K} \frac{1}{\sigma_{X,k}^2} \|x-\mu_{X,k}\|^2 + \frac{1}{\sigma_{Y,k}^2} \|y-\mu_{Y,k}\|^2 \;.$$

That is, supposing gaussian-like classifiers with covariances parallel to the axis, the product rule tries to minimize the sum of the squared distances to the respective "centers" of classes for X and Y, such distances being weighted by the inverse of the "spread" of the the classes (an intuitively reasonable strategy, in fact).

Proof. Under the mentioned hypothesis, we have

$$\begin{gathered} \max\nolimits_{k=1,\ldots,K} p_{X,k}(x) \cdot p_{Y,k}(y) = \\ = \max\nolimits_{k=1,\ldots,K} e^{-\left(\frac{1}{2\sigma_{X,k}^2}\|x-\mu_{X,k}\|^2 + \frac{1}{2\sigma_{Y,k}^2}\|y-\mu_{Y,k}\|^2\right)} \;. \end{gathered}$$

Applying log and multiplying by 2 the second member of the above equality results in

$$\begin{gathered} \max\nolimits_{k=1,\ldots,K} p_{X,k}(x) \cdot p_{Y,k}(y) = \\ = \min\nolimits_{k=1,\ldots,K} \frac{1}{\sigma_{X,k}^2}\|x-\mu_{X,k}\|^2 + \frac{1}{\sigma_{Y,k}^2}\|y-\mu_{Y,k}\|^2 \;. \end{gathered}$$

□

Fact 3. *Let us now define confidence functions as follows:*

$$p_{X,k}(x) = \frac{1}{(2\pi)^{d_X} |\Sigma_{X,k}|^{1/2}} e^{-\frac{1}{2}(x-\mu_{X,k})^\top \Sigma_{X,k}^{-1}(x-\mu_{X,k})} \text{ , and}$$

$$p_{Y,k}(y) = \frac{1}{(2\pi)^{d_Y} |\Sigma_{Y,k}|^{1/2}} e^{-\frac{1}{2}(y-\mu_{Y,k})^\top \Sigma_{Y,k}^{-1}(y-\mu_{Y,k})} \;,$$

where, for each $Z \in \{X,Y\}$, $|\Sigma_{Z,k}|$ is the determinant of $\Sigma_{Z,k}$. Let us suppose also that, conditioned to the class c_j, X and Y are uncorrelated, that is, being Σ_k the covariance of $(X,Y)|C=c_k$, we can write

$$\Sigma_k = \begin{bmatrix} \Sigma_{X,k} & 0 \\ 0 & \Sigma_{Y,k} \end{bmatrix} \;,$$

where, for each $Z \in \{X,Y\}$, $\Sigma_{Z,k}$ is the covariance of $Z|C=c_k$. Then, putting $\mu_j = (\mu_{X,j}, \mu_{Y,j})$, we have

$$\begin{gathered} p_{X,k}(x) \cdot p_{Y,k}(y) = \\ = \frac{1}{(2\pi)^{d_X+d_Y} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}((x,y)-\mu_k)^\top \Sigma_j^{-1}((x,y)-\mu_k)} \;. \end{gathered}$$

That is, supposing gaussian classifiers, the product rule is equivalent to learning using the concatenated vectors of features.

Proof. The inverse of Σ_k is

$$\Sigma_k^{-1} = \begin{bmatrix} \Sigma_{X,k}^{-1} & 0 \\ 0 & \Sigma_{Y,k}^{-1} \end{bmatrix} \;.$$

This way, the expression

$$(x-\mu_{X,k})^\top \Sigma_{X,k}^{-1}(x-\mu_{X,k}) + (y-\mu_{Y,k})^\top \Sigma_{Y,k}^{-1}(y-\mu_{Y,k})$$

reduces to

$$((x,y)-\mu_k)^\top \Sigma_k^{-1}((x,y)-\mu_k) \;.$$

Now

$$\frac{1}{(2\pi)^{d_X}|\Sigma_{X,k}|^{1/2}} \cdot \frac{1}{(2\pi)^{d_Y}|\Sigma_{Y,k}|^{1/2}} = \frac{1}{(2\pi)^{d_X+d_Y}|\Sigma_k|^{1/2}} \ .$$

Therefore

$$p_{X,k}(x) \cdot p_{Y,k}(y) =$$
$$= \frac{1}{(2\pi)^{d_X+d_Y}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}((x,y)-\mu_k)^\top \Sigma_k^{-1}((x,y)-\mu_k)} \ .$$

□

3 Discussion

According to Fact 1, the product rule arises when maximizing the posterior under the hypothesis of equivalent priors and conditional independence given a class. We have just seen (Fact 3) that, supposing only uncorrelation (which is less then independency), the product rule appears as well. But in fact we have used gaussian classifiers, i.e., we supposed the data was normally distributed. This is in accordance with the fact that normality and uncorrelation implies independency.

An important consequence of Fact 3 has to do with the *curse of dimensionality*. If there is strong evidence that the conditional joint distribution of (X,Y) given any class $C = c_k$ is well approximated by a normal distribution, and that $X|C = c_k$ and $Y|C = c_k$ are uncorrelated, than the product rule is an interesting option, because we do not have to deal with a feature vector with dimension larger the largest of the dimensions of the original descriptors. Besides, the product rule allows parallelization.

REFERENCES

[1] L. Alexandre, A. Campilho and M. Kamel. *On Combining Classifiers Using Sum and Product Rules*. Pat. Rec. Letters 22. P. 1283-1289. 2001.

[2] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, Cambridge, MA, 2004.

[3] M. van Breukelen, R. Duin, D. Tax and J. Hartog. *Handwritten Digit Recognition by Combined Classifiers*. Kybernetica, Vol. 34, Number 4, P. 381-386. 1998.

[4] M. Cicconet. *The Guitar as a Human-Computer Interface*. D.Sc. Thesis. National Institute of Pure and Applied Mathematics. Rio de Janeiro, 2010.

[5] M. Cicconet, P. Carvalho and L. Velho. *On Bimodal Guitar-Chord Recognition*. International Computer Music Conference. New York, 2010.

[6] R. Duin and D. Tax. *Experiments with Classifier Combining Rules*. 1st Int. Workshop on Multiple Classifier Systems. P. 16-29. London, UK. 2000.

[7] J. Kittler, M. Hatef, R. Duin and J. Matas. *On Combining Classifiers*. IEEE TPAMI, Vol. 20, N. 3, March 1998.

[8] S. Li and C. Zong. *Classifier Combining Rules Under Independence Assumptions*. 7th International Conference on Multiple Classifier Systems. Springer-Verlag. Berlin Heidelberg. 2007.

[9] T. Mertens, J. Kautz and F. Van Reeth. *Exposure Fusion*. 15th Pacific Conference on Computer Graphics and Applications. P. 382-390. Washington, DC, USA. 2007.